International Journal of Innovative Computing 7(2) 1-5

International Journal of Innovative Computing

Journal Homepage: http://se.fsksm.utm.my/ijic/index.php/ijic

Prediction of Students' Performance in E-Learning Environment Using Random Forest

Yusuf Abubakar[1,2]
[1]Department of Computer Science
Universiti Teknologi Malaysia
Johor Bahru, Malaysia
yusufabukausar@gmail.com

[2]Department of Computer Science
Nuhu Bamalli Polytechnic
Zaria, Nigeria

Nor Bahiah Hj Ahmad
Department of Software Engineering
Universiti Teknologi Malaysia
Johor Bahru, Malaysia
bahiah@utm.my

***Abstract* – The need for advancement in e-learning technology causes educational data to become very huge and increase rapidly. The data is generated on daily basis as a result of students' interaction with learning management systems. The data contains hidden information about participation of students in various activities of e-learning which when revealed can be used to associate with the students' performance. Predicting the performance of students based on the use of e-learning system in educational institutions is a major concern and has become very important for education managements to better understand why so many students perform poorly or even fail in their studies. However, it is difficult to do the prediction due to the diverse factors or characteristics that influence their performance. This paper is aimed at predicting students' performance by considering the students interaction in e-learning environment, their assessment marks and prerequisite knowledge as prediction features. Random Forest algorithm has been used for the prediction. Results show that the algorithm outperforms the popular decision tree and K-Nearest Neighbor algorithms. In addition to the performance prediction, the research findings also revealed most significant attributes that influences students' performance.**

Keywords — E-Learning, Students' Performance, prediction, learning management system

I. INTRODUCTION

Student performance in educational institutions such as Universities and Colleges is not only a pointer to the effectiveness of the institutions but also major determinant of the future of students in particular and nations at large. Learning outcomes have become phenomenon of interest to all and this account for the reason why scholars have been working hard to find out factors that militate against good academic performance [10]. As a result, academic achievement of learners has attracted attention of scholars, parents, policymakers and planners, their goal is to work hard towards attainment of academic excellence by students.

Performance of students may be influenced by several factors such as gender, age, parents' socioeconomic situation, area of resident, nature of school being attended, school medium of teaching, number of study hours spent daily, and nature of accommodation which may be school own hostel or otherwise [12]. A number of researches about factors affecting students' performance at different study levels have been conducted by many authors.

Students' performance prediction is one of the earliest and most valuable applications of Educational Data Mining (EDM) and its objective is to measure the hidden value of students' performance, understanding or grade from the other information, attitude or behavior of those students. This is a

difficult issue to address because of the diverse number of factors that influences the performance of students [12].

Several EDM techniques have been used in the prediction of students' performance such as classification, clustering and association rule. It is essential to note that most recent researches on EDM for students' performance prediction were primarily applied to cases of University and high school students [2] and specifically, in most cases to e-learning or related mode of instruction [12]. This is fundamentally as a result of increase in the use of learning management systems (LMSs).

This paper therefore focuses on developing prediction model of students' academic performance based on their interaction with learning management system in order to explore the performance of Random Forest in the prediction of student performance with the aim of achieving high prediction accuracy.

II. RELATED WORKS

The research in e-learning domain is facilitated by the extensive amount of data stored by the e-learning systems, most of these systems have the ability to collect data about the student activities, tracking navigational pathways through educational resources, time spent on various topics, or number of visits. [2] present the main findings of an educational data mining survey covering the period 1995-2005. [3] made another survey covering the latest data mining approach in education domain. Both surveys show that the number of data mining applications in education is constantly increasing, and they cover a lot of educational processes such as: enrollment management, academic performance, web-based education, retention. Many case studies on data mining techniques in education are cited in the literature, [4][2][5]. These case studies aim at predictions of student performance, mainly through classification techniques in other to relate student related variables with academic performance. [6] proposed a model for the application of data mining in higher education. A model was developed to find similar patterns from the data gathered and to make predication about students' performance [7][8][10]. [9] presented different case studies on educational data mining. One of these studies intended to highlight factors that determine the academic success of first-year students. The methods used are classification and regression trees and neural networks. There were generated decision trees, and association rules. A sensitivity analysis was performed to analyze factors. Variables considered were demographic variables and performance indicators before college. By this analysis can be achieved overall average prediction in the year. The analysis carried out by two classes of methods showed that the most important factors for academic success in first year of college are SAT scores (average of high school equivalent) and position in the rankings achieved on average in high school. [13] used Random Forest to develop a students' performance prediction model with new category of features called behavioral features which relates to the learner interactivity with the e-learning using data obtained from a LMS called Kalboard 360. WEKA data mining tool was used to conduct the experiment and the result obtained shows that there is up to 25.8% accuracy improvement using the Random Forest algorithm. It was also discovered that there is a strong relationship between learners' behavior and their academic achievement.

III. METHOD

A. Random Forest

Random Forest is supervised ensemble machine learning approach for classification, regression and other tasks that operates by constructing a number of decision trees during training and producing as its output the class that is mode of the classes of the individual trees [13]. Unlike in decision tree where each node is split using the best among the attributes, in Random Forest each node is split using the best among a subset of predictors randomly chosen at the node. This strategy makes Random Forest perform very well when compared to many other classification algorithms including Neural Network, Support Vector Machine and Discriminant Analysis among others and it is robust against overfitting.

B. Data Preparation

For the purpose of this research, only the students' interaction data on the e-learning activities, students' prerequisite knowledge and assessment were selected because they provide information on the students' participation in the course activities as well as the impact of prior knowledge and assessment on students' performance. The basis for selecting these attributes is inline with [13] who selected 10 attributes from students' interaction with MOODLE LMS and assessment. Also [1] selected 11 attributes from MOODLE interaction and prerequisite knowledge. For the purpose of this research, the data altogether consist of 26 students' records and a total of eleven (11) attributes were selected including the class attribute, these attributes were obtained from three different set of features. Table I gives a description of the selected attributes.

TABLE I. Attributes Description

SN	Attribute	Description
1.	CourseView	Number of course views during semester
2.	AssignView	Number of assignment views during semester
3.	Assign_sumbit_update	Number of assignment uploads and updates during semester
4.	ResourceView	Number of resource views during semester
5.	ForumView	Number of forum views during semester

6.	PT I	Overall score in programming technique I
7.	PT II	Overall score in programming technique II
8.	Assignment	Total score in all assignments for the semester
9.	LabTotal	Total score in all lab work in Data structure
10.	Midterm	Score in midterm examination
11.	Performance	The students overall grade (High, Medium, Low)

After selecting the target data, it was observed that the dataset was having some missing values and also a column representing the participation of students in discussion forum was having zeros for all students.

IV. EXPERIMENT AND RESULTS

A. Performance Prediction using Random Forest

The dataset after pre-processing was exported to MATLAB software and codes were written to implement the Random Forest algorithm in order to investigate its performance in predicting students' academic performance on the dataset. A 10-fold cross validation was used to train and validate the model after which its performance was measured using confusion matrix. The aim of the experiment is to investigate the performance of Random Forest classifier in carrying out prediction. After conducting the experiment, the Random Forest algorithm obtained a prediction accuracy of 76.9%.

In addition to prediction accuracy, other measure including precision, recall and F-measure were also evaluated for the Random Forest algorithm. Table 2 gives detailed results of the precision, recall and F-measure on Random Forest algorithm.

TABLE II: Random Forest Evaluation Measures

	PRECISION (%)	**RECALL (%)**	**F-MEASURE (%)**
HIGH	82	93	88
LOW	100	50	67
MEDIUM	57	57	57

Table 2 shows the results obtained by Random Forest for precision, recall and F-measure. Precision measure the percentage of tuples that the classifier marked and classified in the class and are actually in the class. The algorithm obtained precision of 82%, 100% and 57% for "High", "Medium" and "Low" class respectively. This shows that Random Forest can accurately identify all students that with "Low" performance since it obtained 100% precision for that class. It also identifies 85% of students with "High" performance, but for the students with "Medium" performance, the algorithm only identified 57%. With respect to recall, the algorithm obtained 93%, 50% and 57% for "High", "Low" and "Medium" classes respectively. Lastly, the algorithm also obtained the F-measure of 88%, 67% and 57% for "High", "Low" and "Medium" classes respectively. Figure I illustrates the Random Forest result for precision, recall and F-measure.

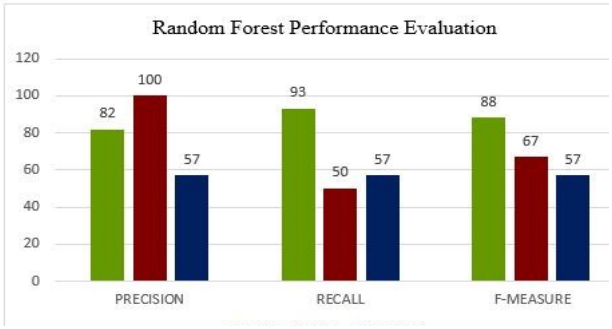

Figure I. Random Forest Evaluation Metrics

B. Generating Significant Attributes using Random Forest

The MATLAB TreeBagger function was used to create the ensemble of Random Trees. 10-fold cross validation was used in training and testing the model. The number of trees parameter is set to 10 because this is the standard used by most researchers, thus this generates a set of 10 random trees. The root node for each tree indicate the most significant attribute for that tree, it is through the root nodes that other internal nodes can be reached in order of significance. After conducting the experiment, the Random Forest algorithm generates 10 random trees in which three (3) of them has LabTotal as a root node, another three (3) has and Assign_sumbit_update as their root note, Midterm and AssignmentView has 2 and 1 trees as their root nodes respectively.

After generating the trees, Random Forest considers LabTotal, Assign_sumbit_update, Midterm and AssignView as the most significant attributes to build the trees since the root node for each of the trees begins with one of these attributes. The random trees obtained by Random Forest algorithm are presented next

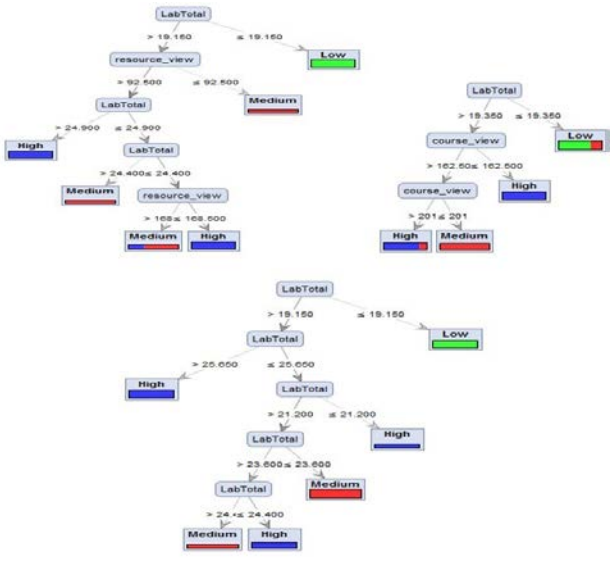

Figure II: Trees with LabTotal as Root Node

Figure II shows the three (3) random trees generated by the Random Forest algorithm having LabTotal as their root node, this signifies that through LabTotal and possibly other attributes, students' performance can be determined. The same types of trees were also created for the other significant attributes. After creating the trees, the algorithm uses a majority voting in deciding which class an instance belongs to. This majority voting strategy selects the class that was predicted by the majority of tree models. Other trees that were generated based on the most significant attributes are presented in Figure III. and Figure IV

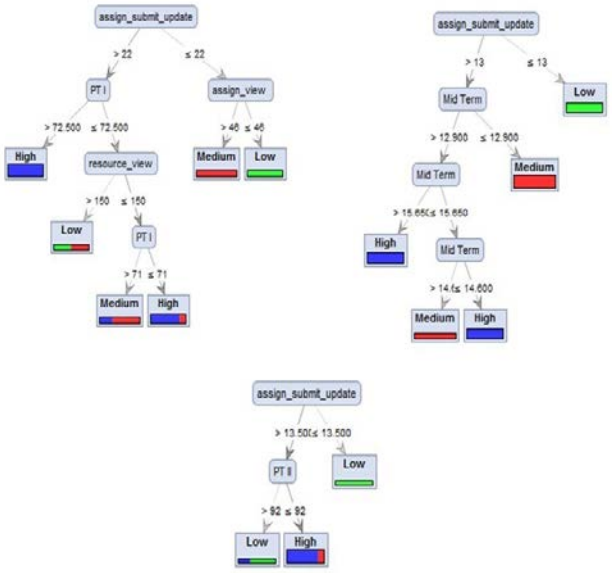

Figure III: Trees with Assign_Submit as Root Node

Figure 4: Trees with Midterm as Root Node

C. Experimental Result with other Techniques

Results of the three classification algorithms (base classifiers) tested are presented in this section. Prediction models using the base classifiers were trained and validated using 10-fold cross validation for the training and testing. The result obtained by the base classifiers will later be compared with the result obtained by Random Forest in other to determine the algorithm that perform better in predicting students' performance. The algorithms used are Naive Bayes, K-Nearest Neighbour and Decision Tree which are the commonly used classifiers among many authors in predicting students' performance as reported by [13]

The results of Random Forest and three non-ensemble techniques (Naive Bayes, k-Nearest Neighbor and Decision Tree) in prediction of students performance are compared. The overall performance of all the techniques is described in terms of Accuracy, Precision, Recall and F-Measure. The accuracy is regarded as the overall correctness of the model, the precision which is the measure of the accuracy provided that a specific class has been predicted, the recall or sensitivity which is a measure of the ability of the prediction model to select instances for a certain class from a dataset and finally the F-measure which is the accuracy of harmonic mean of precision and recall. Table III shows the accuracies of the algorithms.

Table III. Classifier Accuracies

Classifier	**Accuracy (%)**
Naive Bayes	92.3
k-Nearest Neighbor	69.2
Decision Tree	61.5
Random Forest	76.9

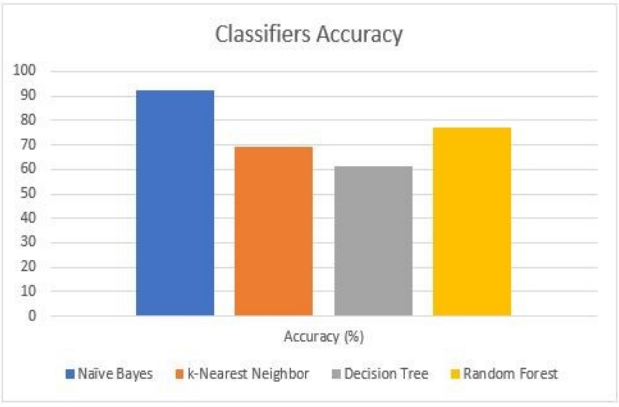

Figure V: Classifiers Accuracy

Figure V depicts the accuracies of four classifiers Random Forest, Naive Bayes, K-Nearest Neighbour and Decision Tree. Even though Random Forest performs reasonably better than other classifiers (k-Nearest Neighbour, and Decision Tree) but the Naive Bayes has outperformed the ensemble in terms of the prediction accuracy. Naive Bayes classifier outperformed all other classifiers with the accuracy of 93.3%, followed by Random Forest with 76.9% and KNN and Decision tree with 69.2% and 61.5% accuracies respectively. The overall performance of a model does not solely depend on the accuracy, other evaluation metrics such as Precision, Recall and F-measure are also criteria that defines the suitability of models especially when there need to evaluate the models in terms of class predictions.

V. DISCUSSION

The experiments conducted revealed the performance of Random Forest compared to other classifiers that have been commonly used in the area of students' performance prediction. Although the Random Forest algorithm performed better than KNN and Decision Tree but it was outperformed by Naive Bayes classifier in prediction accuracy. Even though Naive Bayes outperformed Random Forest, however, Random Forest presents more information than Naive Bayes in indicating the attributes that are more important in predicting students' performance. Random Forest obtained accuracy of 76.9% which was better than KNN and Decision Tree that recorded the accuracy of 69.2% and 61.5% respectively. In the case of precision, recall and f-measure, Random Forest also performed better than KNN and Decision Tree but was again outperformed by Naive Bayes.

REFERENCES

[1] Sisovic, S., Matetic, M. and Bakaric, M. B. (2015). Mining student data to assess the impact of moodle activities and prior knowledge on programming course success. In *Proceedings of the 16th International Conference on Computer Systems and Technologies*. ACM, 366–373.

[2] Kotsiantis, S., Patriarcheas, K. and Xenos, M. (2010). A combinational incremental ensemble of classifiers as a technique for predicting students performance in distance education. *Knowledge-Based Systems*. 23(6), 529–535.

[3] Liang, J., Yang, J., Wu, Y., Li, C. and Zheng, L. (2016). Big Data Application in Education: Dropout Prediction in Edx MOOCs. In *Multimedia Big Data (BigMM), 2016 IEEE Second International Conference on*. IEEE, 440–443.

[4] Mining VRSEC student learning behaviour in moodle system using datamining techniques. In *Computer and Communications Technologies (ICCCT), 2014 International Conference on*. IEEE, 1–7.

[5] Namdeo, J. and Jayakumar, N. (20^{14}). Predicting Students Performance Using Data Mining Technique with Rough Set Theory Concepts. *International Journal*. 2(2).

[6] Opitz, D. and Maclin, R. (1999). Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research*. 11, 169–198.

[7] Pandey, U. K. and Pal, S. (2011). Data Mining: A prediction of performer or underperformer using classification. *arXiv preprint arXiv:1104.4163*.

[8] Panigrahi, S., Kundu, A., Sural, S. and Majumdar, A. K. (2009). Credit card fraud detection: A fusion approach using Dempster–Shafer theory and Bayesian learning. *Information Fusion*. 10(4), 354–363.

[9] Aremu, A. and Sokan, B. (2003). A multi causal evaluation of academic performance of Nigerian learners: Issues and implications for national development. *Department of Guidance and counseling, University of Ibadan.*

[10] Pyle, D. (2005). Data Preparation and Preprocessing. In *Do Smart Adaptive Systems Exist?* (pp. 27–53). Springer.

[11] Romero, C., Lopez, M.-I., Luna, J.-M. and Ventura, S. (2013). Predicting students'´ final performance from participation in on-line discussion forums. *Computers & Education*. 68, 458–472.

[12] Romero, C., Lopez, M.-I., Luna, J.-M. and Ventura, S. (2013). Predicting students'´ final performance from participation in on-line discussion forums. *Computers & Education*. 68, 458–472.

[13] Romero, C. and Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert systems with applications*. 33(1), 135–146.